Journal of Multivariate Analysis 109 (2012) 221–235

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Non-convex penalized estimation in high-dimensional models with single-index structure

Tao Wang, Pei-Rong Xu, Li-Xing Zhu *

Hong Kong Baptist University, Hong Kong, China
East China Normal University, China

ARTICLE INFO

Article history:
Received 27 October 2011
Available online 29 March 2012

AMS 2000 subject classifications:
62H12
62G20

Keywords:
High-dimensional variable selection
Minimax concave penalty
Oracle property
Penalized least squares
SCAD
Single-index model

ABSTRACT

As promising alternatives to the LASSO, non-convex penalized methods, such as the SCAD and the minimax concave penalty method, produce asymptotically unbiased shrinkage estimates. By adopting non-convex penalties, in this paper we investigate uniformly variable selection and shrinkage estimation for several parametric and semi-parametric models with single-index structure. The new method does not need to estimate the involved nonparametric transformation or link function. The resulting estimators enjoy the oracle property even in the "large p, small n" scenario. The theoretical results for linear models are in parallel extended to general single-index models with no distribution constraint for the error at the cost of mild conditions on the predictors. Simulation studies are carried out to examine the performance of the proposed method and a real data analysis is also presented for illustration.

1. Introduction

Variable selection is a fundamental task for statistical modeling in high-dimensional settings, where the number of predictors is often comparable to, or even much larger than the total sample size. Traditional variable selection procedures follow either best subset selection or its stepwise variants. However, subset selection is computationally prohibitive when the number of predictors is large. Moreover, as analyzed by Breiman [6], subset selection may suffer from instability because of its inherent discreteness. To deal with these drawbacks, various penalized methods have been proposed during the past years to perform variable selection and shrinkage estimation simultaneously. In particular, the LASSO [36] and the SCAD [16] are two very popular methods with promising computational and statistical properties.

There is a huge literature devoted to studying the theoretical properties of the LASSO, particularly in the linear regression context. See, for instance, [25,15,42,41,3], among many others. Despite its popularity, the LASSO does suffer from several drawbacks, the most severe of which is its estimation bias. To this end, Fan and Li [16] proposed the SCAD in a general parametric framework. When the number of predictors is finite, they studied the oracle properties of general non-concave penalized likelihood estimators. Here, the oracle property means that the estimator is asymptotically as efficient as the ideal one assisted by an oracle who knows which coefficients are nonzero and which are zero. Their results were later extended by Fan and Peng [18] to the setting with a diverging number of predictors. Recently, Kim et al. [24] proved that for linear models, the oracle property of the SCAD continues to hold while the number of predictors can grow at a polynomial rate, up to exponentially fast, of the sample size. Other works on the advantages of penalized methods with non-convex penalties over the LASSO include [30,40]. In particular, Zhang [40] investigated in detail the properties of the minimax concave penalty

* Correspondence to: Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, China.
E-mail address: lzhu@hkbu.edu.hk (L.-X. Zhu).

0047-259X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2012.03.009

approach. Fan and Lv [17] gave a selective overview on the theoretical properties as well as algorithmic implementations of penalized likelihood methods in high-dimensional settings. We mention that alternative sparsity promoting approaches, such as the PAC-Bayesian approach using sparsity favoring priors, have also been proposed and well-studied [12,13].

In most empirical applications of regression analysis, however, the working model such as the linear regression model is at best an approximation. Because of the so-called "curse of dimensionality", it is very difficult or even infeasible to formulate and then validate a parametric model with a large number of predictors. To mitigate the risk of model misspecification and to overcome the curse of dimensionality, semiparametric models have attracted much attention in the literature. Popular models include the response transformation model $g_1(Y) = \boldsymbol{\beta}^T X + \epsilon$ and the classical single-index model $Y = g_2(\boldsymbol{\beta}^T X) + \epsilon$. Here $g_1(\cdot)$ is an (unknown) monotone function, $g_2(\cdot)$ is an unknown link function, and ϵ is assumed to be independent of $X = (X_1, \ldots, X_p)^T$. See [21,22] for more details. Interestingly, these two classes of models are of a common feature in model structure: other than an unknown nonparametric model transformation or link function, the information of the response can be captured through a single linear combination of the predictors. We call this the single-index structure. In this paper, we consider the following class of models with the single-index structure

$$Y = g(\boldsymbol{\beta}^T X, \epsilon), \tag{1.1}$$

or equivalently

$$Y \perp\!\!\!\perp X | \boldsymbol{\beta}^T X, \tag{1.2}$$

where g is an unspecified bivariate function and $\perp\!\!\!\perp$ indicates independence. The statement is thus that, given $\boldsymbol{\beta}^T X$, the response variable Y and the predictor vector X are independent of each other. Many important regression models, including linear models and generalized linear models, naturally satisfy (1.2). Other examples are the transformation linear model and the classical single-index model mentioned above.

The family of general single-index models (1.2) have been well-studied in the literature. On one hand, promising methods for estimating the index $\boldsymbol{\beta}$ include least squares method [29], structural adaptation method [23,11], and those in the sufficient dimension reduction context, such as sliced inverse regression method [28], sliced average variance estimation method [10], directional regression method [27], discretization–expectation estimation method [43], and many others. On the other hand, some attempts have also been made to address the variable selection problem. Kong and Xia [26] and Naik and Tsai [31] proposed new selection criteria for variable selection in the classical single-index model. *See also* [2]. Other alternatives are model free, which typically integrate sufficient dimension reduction techniques with the regularization paradigm, see [4,39] and the references therein. In particular, Wu and Li [39] investigated the asymptotic properties of sufficient dimension reduction estimators equipped with a SCAD-type penalty, when the number of predictors diverges to infinity with the sample size. The approaches and results, however, cannot be directly extended to the "$p > n$" setting. Therefore, it is of great interest to see whether the model-based selection methods, such as those in [24,40], have their justifiable counterparts in the general setting of (1.2) where no parametric model is imposed.

In this article, we investigate index estimation and variable selection, with an emphasis on the latter, for the class of models (1.2) with high-dimensional predictors. First, we study the asymptotic properties of index estimation. For any bounded transformation of the response, we propose an index estimator and establish the consistency and asymptotic normality in the presence of a diverging number of predictors. Second, we briefly discuss the choice of response transformation and adopt a response–distribution transformation considered in [38]. Third, by introducing a non-convex penalty function, we consider the penalized least squares optimization. We prove the oracle property of the SCAD and the minimax concave penalty estimator, while we allow the number of predictors to grow at some polynomial rate of the sample size. Finally, we evaluate the finite sample performance of the proposed method through simulation studies as well as a real data analysis. All technical proofs are given in the Appendix.

2. Methodology and main results

2.1. Index estimation and asymptotics

We adopt the least squares approach to estimate the index $\boldsymbol{\beta}$. It is very simple to use. In addition, the proposed least squares estimation allows us to directly introduce the penalty function, as given in Section 2.3. Of course, before developing any justifiable variable selection procedure, it is important to establish the asymptotic properties for the unpenalized estimation. We shall address this problem in this subsection.

Let $\boldsymbol{\Sigma} = \mathrm{Cov}(X)$ and $\boldsymbol{\sigma} = \mathrm{Cov}\{X, h(Y)\}$ for a given function $h(\cdot)$ of the response. We assume that $\boldsymbol{\Sigma}$ is positive definite. Define $\boldsymbol{\beta}_h = \boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}$ as the coefficient vector of the least squares type. The following proposition follows immediately from Theorem 2.1 in [29].

Proposition 1. *Assume that $E(X|\boldsymbol{\beta}^T X)$ is a linear function of $\boldsymbol{\beta}^T X$. Then $\boldsymbol{\beta}_h$ is proportional to $\boldsymbol{\beta}$, that is, $\boldsymbol{\beta}_h = \kappa_h \times \boldsymbol{\beta}$ for some constant κ_h.*

The design condition of Proposition 1, known as the linearity condition, is satisfied when X has an elliptical distribution. It is widely assumed in the sufficient dimension reduction literature, see [28,8,9], among others. Hall and Li [20] proved that, as p tends to infinity, such a linearity can hold to a reasonable approximation in many problems. Proposition 1 indicates that,

under the linearity condition, the index can be estimated without employing any nonparametric techniques. This is indeed a trade-off between estimation efficiency and technical conditions. However, the empirical studies, including those reported in Section 3, show that the performance of our method is robust against the violation of this linearity. Finally, we note that there is no guarantee that the constant κ_h is different from zero. For example, consider the noiseless model: $Y = (\boldsymbol{\beta}^T X)^2$. It follows that $\boldsymbol{\beta}_h$ is simply the null vector for any h. The propose method is applicable to situations where there is some linear trend in the regression, such as $Y = (\boldsymbol{\beta}^T X + c)^2$ for $c \neq 0$.

Let $\{(\mathbf{x}_i, y_i) : \mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T\}_{i=1}^n$ be a random sample on (X, Y). We denote by $\mathbf{y} = (y_1, \ldots, y_n)^T$ the response vector and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ the design matrix with jth column $\mathbf{X}_j = (x_{1j}, \ldots, x_{nj})^T$. Let $\mathbf{h}(\mathbf{y}) = (h(y_1), \ldots, h(y_n))^T$. Without loss of generality, we assume that both $\mathbf{X}$ and $\mathbf{h}(\mathbf{y})$ are centered, so that the intercept is not included in the regression function. By Proposition 1, we define the least squares index estimator of $\boldsymbol{\beta}_h$ as

$$\hat{\boldsymbol{\beta}}_h = \arg\min_{\mathbf{b}} \|\mathbf{h}(\mathbf{y}) - \mathbf{X}\mathbf{b}\|_2^2 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{h}(\mathbf{y}). \tag{2.1}$$

Remark 1. The sole reason why we call this estimator the least squares is because of its resemblance in form to the ordinary least squares for linear models. However, if we write $h(Y) = \boldsymbol{\beta}_h^T X + e$, then the "error" e must satisfy $E(eX) = 0$ but with no conclusion further that $E(e|X) = 0$. As a result, neither the residual-based methods nor the likelihood approaches are applicable unless we consider nonparametric plug-in estimation for the nonparametric function. As is well known, such a way is however not an effective way of estimation and computation. We shall show in this and later subsections that, by exploiting this simple structure, it is possible to estimate the index and select significant predictors without imposing any parametric model structure. Actually, what makes the things work is the fact that Y is independent of X given $\boldsymbol{\beta}^T X$.

In the rest of this subsection, we investigate the asymptotic properties of $\hat{\boldsymbol{\beta}}_h$ when the number of predictors p diverges to infinity with the sample size n. We note that, hereafter, most of quantities and data objects are functions of n, but this dependence on n is often left implicit, especially for n-vectors and matrices with n rows.

Let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ be the smallest and largest eigenvalues of a symmetric matrix, respectively. We need the following technical conditions:

(A1) $0 < L_1 < \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) < L_2 < \infty$ for some L_1 and L_2,
(A2) $\max_{1\leq j\leq p} E(X_j^4) < L_3 < \infty$ for some L_3,
(A3) $p = o(n^{1/2})$.

Conditions (A1) and (A3) are quite reasonable and are widely assumed in the high-dimensional literature [18,44]. Condition (A2) is a standard moment condition, which is typically needed even in the fixed-dimensional setting. We do not need any condition on the model error.

Let $\tilde{h}(y) = h(y) - E\{h(Y)\}$ and $\boldsymbol{\Lambda} = \mathrm{Cov}[\{\tilde{h}(Y) - X^T\boldsymbol{\beta}_h\}X]$. The consistency and asymptotic normality are stated in the following two theorems.

Theorem 1. *Assume the conditions* (A1)–(A3). *For any bounded transformation function* $h(\cdot)$, *we have* $\|\hat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}_h\|_2 = O_P(p/\sqrt{n})$.

Theorem 2. *Assume the conditions of Theorem 1. Further assume that*

$$\max_{1\leq j\leq p} E(X_j^8) < L_4 < \infty \quad \text{and} \quad p = o(n^{1/4}).$$

For any vector $\mathbf{v} \in \mathbf{R}^p$ *such that* $\|\mathbf{v}\|_2 \leq 1$ *and* $\mathbf{v}^T\boldsymbol{\Lambda}\mathbf{v} \to G > 0$ *as* $n \to \infty$, *we have*

$$\sqrt{n}\mathbf{v}^T(\hat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}_h) \xrightarrow{L} N(0, G), \quad \text{as } n \to \infty.$$

Theorems 1 and 2 indicate that $p = o(n^{1/2})$ and $p = o(n^{1/4})$ are required to obtain the estimation consistency and the asymptotic normality, respectively. These results are of interest on their own. Note that the rates are not optimal in the sense that, under additional complicated conditions, we can achieve an improvement in rates by applying much more sophisticated techniques such as those in [33].

2.2. A distribution transformation

The above results hold for any bounded transformation function h. In this subsection, we propose a response–distribution transformation. Let F_Y be the marginal distribution function of Y. By Proposition 1, $\boldsymbol{\beta}_{F_Y}$ is proportional to the index $\boldsymbol{\beta}$. Thus, one can use as function h the distribution function F_Y of the response. See also [38].

Because the distribution F_Y is unknown in practice, we may use the empirical distribution F_n, given by $F_n(y) = n^{-1}\sum_{i=1}^n \mathbf{1}_{\{y_i\leq y\}}$. Let $\mathbf{F}_n(\mathbf{y}) = (F_n(y_1), \ldots, F_n(y_n))^T$. We define the distribution-transformation least squares estimator as

$$\hat{\boldsymbol{\beta}}_{F_n} = \arg\min_{\mathbf{b}} \|\mathbf{F}_n(\mathbf{y}) - 1/2 - \mathbf{X}\mathbf{b}\|_2^2 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\{\mathbf{F}_n(\mathbf{y}) - 1/2\}. \tag{2.2}$$

We have derived the asymptotic properties of $\boldsymbol{\beta}_{F_Y}$ in Section 2.1. By the law of the iterated logarithm for empirical distributions, the same results hold for $\hat{\boldsymbol{\beta}}_{F_n}$ as well. The proof is simple and is omitted here.

Empirical studies show that the response–distribution transformation performs well. If we can estimate the optimal transformation efficiently, the performance might be even better. Nevertheless, how to estimate such an optimal function is an important yet challenging question for future study. When there are a large number of predictors, it is reasonable to expect some of them to be irrelevant. If we can select important predictors consistently, however, we may achieve a substantial improvement of performance by applying more efficient estimation procedures to the identified set of predictors. The variable selection problem is addressed in the next subsection.

2.3. Model-free variable selection in high dimensions

Building upon the least-squares index formulation and the distribution-based response transformation, we consider the following penalized least squares problem

$$Q_\lambda(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}\{F_n(y_i) - 1/2 - \mathbf{x}_i^T\boldsymbol{\beta}\}^2 + \sum_{j=1}^{p} J_\lambda(|\beta_j|), \tag{2.3}$$

where $J_\lambda(\cdot)$ is a penalty function and $\lambda > 0$ is a regularization parameter.

As advocated by Fan and Li [16], any good penalty function should give estimators with three desired properties, including asymptotic unbiasedness, sparsity and continuity. The L_1 penalty $J_\lambda(|t|) = \lambda|t|$, however, does not enjoy the unbiasedness. Fan and Li [16] suggested using the SCAD penalty, defined as $J_\lambda(0) = 0$ and

$$J'_\lambda(|t|) = \lambda\left\{\mathbf{1}_{\{|t|\le\lambda\}} + \frac{(a\lambda - |t|)_+}{(a-1)\lambda}\mathbf{1}_{\{|t|>\lambda\}}\right\} \quad \text{for some } a > 2. \tag{2.4}$$

Often $a = 3.7$ is recommended. Recently, Zhang [40] studied the penalized least squares estimation in high-dimensional linear models and proposed the minimax concave penalty which is given by $J_\lambda(0) = 0$ and

$$J'_\lambda(|t|) = \lambda\left(1 - \frac{|t|}{a\lambda}\right)\mathbf{1}_{\{|t|\le a\lambda\}} \quad \text{for some } a > 1. \tag{2.5}$$

Both the SCAD penalty and the minimax concave penalty are non-convex and satisfy the aforementioned three properties simultaneously. We consider below the optimization (2.3) with the SCAD penalty and the minimax concave penalty.

Suppose that the true vector $\boldsymbol{\beta}_0 \equiv \boldsymbol{\beta}_{F_Y} = (\beta_{0,1}, \ldots, \beta_{0,p})^T$ is sparse. Without loss of generality, we assume that $\beta_{0j} \neq 0$ for $j = 1, \ldots, q$ and $\beta_{0j} = 0$ for $j = q+1, \ldots, p$. Let $\boldsymbol{\beta}_{(1)} = (\beta_{0,1}, \ldots, \beta_{0,q})^T$ and $\boldsymbol{\beta}_{(2)} = (\beta_{0,q+1}, \ldots, \beta_{0,p})^T$. Let $X_{(1)} = (X_1, \ldots, X_q)^T$ and $X_{(2)} = (X_{q+1}, \ldots, X_p)^T$. We write $\mathbf{X} = (\mathbf{X}_{(1)}, \mathbf{X}_{(2)})$ with $\mathbf{X}_{(1)}$ being a submatrix formed by the first q columns of $\mathbf{X}$ and $\mathbf{X}_{(2)}$ being formed by the last $p-q$ columns. Define $\boldsymbol{\Sigma}_{(i,j)} = \mathrm{Cov}(X_{(i)}, X_{(j)})$ and $\mathbf{C}_{(i,j)} = n^{-1}\mathbf{X}_{(i)}^T\mathbf{X}_{(j)}$ for $i, j = 1, 2$.

Let $\hat{\boldsymbol{\beta}}^o_{(1)}$ be the minimizer of $\|\mathbf{F}_n(\mathbf{y}) - 1/2 - \mathbf{X}_{(1)}\mathbf{b}\|_2^2$. The asymptotic properties of $\hat{\boldsymbol{\beta}}^o_{(1)}$ have been established in Sections 2.1 and 2.2. Define $\hat{\boldsymbol{\beta}}^o = (\hat{\boldsymbol{\beta}}^{oT}_{(1)}, \mathbf{0}^T_{(2)})^T$ as the least squares oracle estimator, where $\mathbf{0}_{(2)}$ is a $(p-q)$-dimensional vector of zeros. We show below that this oracle estimator is asymptotically a local minimum of $Q_\lambda(\boldsymbol{\beta})$. The following regularity conditions are needed.

(A) $\lambda_{\min}(\boldsymbol{\Sigma}_{(1,1)}) \ge M_1$ for some $M_1 > 0$,
(B) $q = O(n^{c_1})$ for some $0 \le c_1 < 1/2$,
(C) $n^{(1-c_2)/2}\min_{1\le j\le q}|\beta_{0,j}| \ge M_2$ for some $c_1 < c_2 \le 1$ and $M_2 > 0$,
(D) $E(X_{(1)}^T\boldsymbol{\beta}_{(1)})^{2k} = O(\|\boldsymbol{\beta}_{(1)}\|_2^{2k})$ for some $k > 0$,
(E) $p = o(n^{(c_2-c_1)k})$, and
(F) $E|X_j - E(X_j)|^{4(c_2-c_1)k+4+\delta} < M_3$ for all $j = 1, \ldots, p$ and some $\delta > 0$.

Condition (D) is satisfied when X follows an elliptically contoured distribution. The moment condition (F) ensures that, under condition (E), the sample covariance matrix $\mathbf{C} = n^{-1}\mathbf{X}^T\mathbf{X}$ converges in probability to $\boldsymbol{\Sigma}$, see Lemma 1 in the Appendix for details. Given that $\lambda_{\min}(\mathbf{C}_{(1,1)})$ is bounded from below almost surely, condition (B) can be weakened to the more natural one: $q = O(n^{c_1})$ for some $0 \le c_1 < 1$. In the case of a deterministic design, Kim et al. [24] studied the oracle properties of the SCAD for high-dimensional linear models. See [40] for the minimax concave penalty approach. The major differences between our work and theirs are as follows. We consider uniformly a family of models with the single-index structure; the new proposed procedure is model free; there is no condition on the model error ϵ, and all the conditions are assumed for the random predictor vector X.

Theorem 3. *Assume the technical conditions* (A)–(F). *For any $\lambda > 0$, let $\mathcal{A}_\lambda$ be the set of local minima of $Q_\lambda(\boldsymbol{\beta})$ with either the SCAD penalty or the minimax concave penalty. If $p = o(\lambda^{2k}n^k)$ and $\lambda = o(n^{-(1-c_2+c_1)/2})$, then $\lim_{n\to\infty} P(\hat{\boldsymbol{\beta}}^o \in \mathcal{A}_\lambda) = 1$.*

Theorem 3 demonstrates that, in probability, the oracle estimator becomes a local minimizer. This property is in fact stronger than the oracle property [16] as it indicates that the resulting estimator is the oracle estimator itself rather than just mimicking the performance of the oracle estimator. If all moments of $X_{(1)}^T\beta_{(1)}$ exist then, by condition (E), p can grow at any polynomial rate. In particular, for Gaussian random designs, we have the following result:

Theorem 4. *Suppose that the design matrix* $\mathbf{X}$ *has independent and identically distributed* $N(0, \boldsymbol{\Sigma})$ *rows. Assume the technical conditions* (A)–(F). *If* $\lambda_n = O(n^{-(1-c_4)/2})$ *and* $p = O(e^{n^{c_3}})$ *with* $0 < c_3 < c_4 < c_2 - c_1$, *then* $\lim_{n\to\infty} P(\hat{\boldsymbol{\beta}}^o \in \mathcal{A}_\lambda) = 1$.

It is an encouraging result that, even when no parametric model is assumed, we have the oracle property while p is allowed to grow much faster than n, up to exponentially fast. In this way, model-free variable selection with justifiable properties is achieved in "$p > n$" situations.

3. Numerical studies

In this section, we aim to evaluate the finite sample performance of the proposed non-convex penalized methods in terms of predictor selectivity and estimation accuracy. We apply the distribution-transformation penalized least squares estimation with the SCAD penalty (D-SCAD) as well as the minimax concave penalty (D-MCP). For comparison purpose, we also examine the distribution-transformation LASSO with the L_1 penalty (D-LASSO). We compute the D-SCAD, the D-MCP and the D-LASSO using the fast and efficient coordinate descent algorithms which are implemented through the publicly available R packages (http://www.r-project.org). To be specific, we use the open-source R packages *glmnet* and *ncvreg*. The former was developed by Friedman et al. [19] who were among the first to advocate the use of coordinate descent algorithms for the LASSO-type estimation, while the latter was recently proposed by Breheny and Huang [5] who examined the application of coordinate descent algorithms for non-convex penalized regression problems. The details of those computational algorithms are then not discussed here. For each competitor, we select its regularization parameter λ by tenfold cross-validation which is a popular and well-established method in the literature. In the D-SCAD and the D-MCP, we need to cross-validate on a two-dimensional surface. To this end, we first pick a relatively small grid of values for a, say (2.2, 2.7, 3.2, 3.7, 4.2) for the SCAD penalty and (1.2, 1.7, 2.2, 2.7, 3.2) for the minimax concave penalty. Then, for each a, we run the algorithm and select λ by tenfold cross-validation. The chosen a is the one giving the smallest cross-validation error.

We consider four models from (1.2) in our simulations:

$$Y = X^T\boldsymbol{\beta}/3 + \epsilon, \tag{3.1}$$

$$Y = |X^T\boldsymbol{\beta} + 1| \times (X^T\boldsymbol{\beta}) + \epsilon, \tag{3.2}$$

$$Y = \text{Poisson}\{\phi(X^T\beta)\}, \tag{3.3}$$

$$Y = \exp(X^T\beta + \epsilon), \tag{3.4}$$

where $\phi(u) = \exp(u)$ is the link function and $\epsilon \sim N(0, 1)$ is independent of X. Thus, the first model is a linear model, the second one has a non-linear structure, the third one belongs to generalized linear models, and the last one is a linear transformation model. Six examples with "$p > n$" are presented here, reflecting different scenarios.

Example 1. We set $n = 100$, $p = 500$ and $q = 5$. The 1st, 3rd, 5th, 7th and 9th components of $\boldsymbol{\beta}$ are given by 3, 1.5, 2, 2 and -2, respectively. The rest are fixed to be zero. The predictors $X_1, \ldots, X_p$ are generated independently from $N(0, 1)$.

Example 2. The same as Example 1, except that the predictors are generated independently from the uniform distribution on the interval $(-\sqrt{3}, \sqrt{3})$.

Example 3. The same as Example 1, except that the predictor vector X is generated from a multivariate normal distribution whose marginal distributions are standard $N(0, 1)$. In addition, the pairwise correlation between the ith and the jth components of X is $0.5^{|i-j|}$ for $i, j = 1, \ldots, p$.

Example 4. The same as Example 2, except that we further multiply X by $\Sigma^{1/2}$ with Σ being the same covariance matrix in Example 3.

Example 5. The same as Example 1, except that $p = 2000$.

Example 6. The same as Example 3, except that $p = 2000$.

Examples 2 and 4 represent the cases where the linearity condition or ellipticity is violated. In Examples 1–4, it is also interesting to compare our methods with existing standard approaches: the SCAD, the minimax concave penalty estimator (MCP) and the LASSO.

To evaluate the empirical performance of $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$, we employ five commonly used measures in the literature: (1) the average model size, MS $= \sum_{j=1}^p \mathbf{1}_{\{\hat{\beta}_j \neq 0\}}$, (2) the average true positive rate, TPR $= q^{-1}\sum_{j=1}^p \mathbf{1}_{\{\hat{\beta}_j \neq 0,\ \beta_j \neq 0\}}$, (3) the average false discovery rate, FDR $= (\sum_{j=1}^p \mathbf{1}_{\{\hat{\beta}_j \neq 0\}})^{-1}\sum_{j=1}^p \mathbf{1}_{\{\hat{\beta}_j \neq 0, \beta_j = 0\}}$, (4) the average squared multiple correlation

Table 1
Simulation results for Example 1.

Selection method	MS (mean)	TPR (mean)	FDR (mean)	COR_1 (mean ± sd)	COR_2 (mean ± sd)
Model (3.1)					
D-SCAD	21.850	0.991	0.7238	0.9439 ± 0.0499	0.9386 ± 0.0535
D-MCP	8.080	0.967	0.2777	0.9534 ± 0.0443	0.9496 ± 0.0474
D-LASSO	31.350	0.987	0.7968	0.9029 ± 0.0550	0.8926 ± 0.0616
SCAD	21.975	0.995	0.7255	0.9611 ± 0.0359	0.9565 ± 0.0403
MCP	7.685	0.976	0.2390	0.9658 ± 0.0376	0.9618 ± 0.0414
LASSO	34.905	0.994	0.8135	0.9174 ± 0.0545	0.9070 ± 0.0636
Model (3.2)					
D-SCAD	11.930	1	0.3566	0.9952 ± 0.0037	0.9946 ± 0.0042
D-MCP	7.010	1	0.1618	0.9949 ± 0.0046	0.9943 ± 0.0052
D-LASSO	34.945	1	0.8126	0.9877 ± 0.0081	0.9862 ± 0.0096
SCAD	13.995	1	0.4717	0.9902 ± 0.0088	0.9892 ± 0.0098
MCP	6.870	1	0.1762	0.9906 ± 0.0095	0.9896 ± 0.0109
LASSO	36.465	1	0.8321	0.9684 ± 0.0213	0.9647 ± 0.0250
Model (3.3)					
D-SCAD	19.755	0.998	0.6585	0.9746 ± 0.0215	0.9712 ± 0.0246
D-MCP	7.405	0.992	0.1991	0.9775 ± 0.0278	0.9754 ± 0.0286
D-LASSO	34.500	0.996	0.8105	0.9352 ± 0.0390	0.9270 ± 0.0459
SCAD	9.355	0.371	0.5164	0.4158 ± 0.2921	0.4067 ± 0.2835
MCP	3.170	0.263	0.2706	0.3886 ± 0.3209	0.3774 ± 0.3119
LASSO	10.720	0.362	0.4537	0.3990 ± 0.2911	0.3854 ± 0.2847
Model (3.4)					
D-SCAD	16.510	1.000	0.5733	0.9841 ± 0.0150	0.9822 ± 0.0173
D-MCP	7.185	1.000	0.1877	0.9858 ± 0.0159	0.9844 ± 0.0181
D-LASSO	35.135	1.000	0.8123	0.9537 ± 0.0289	0.9478 ± 0.0346
SCAD	7.105	0.256	0.4848	0.2906 ± 0.2684	0.2841 ± 0.2518
MCP	1.910	0.145	0.2209	0.2339 ± 0.2849	0.2264 ± 0.2752
LASSO	7.740	0.226	0.3777	0.2521 ± 0.2647	0.2429 ± 0.2520

coefficient, $COR_1 = (\hat{\boldsymbol{\beta}}^T \boldsymbol{\Sigma} \hat{\boldsymbol{\beta}})^{-1} (\hat{\boldsymbol{\beta}}^T \boldsymbol{\Sigma} \boldsymbol{\beta})^2 (\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta})^{-1}$, and (5) the average absolute vector correlation, $COR_2 = |\mathrm{Corr}(\mathbf{X}\hat{\boldsymbol{\beta}}, \mathbf{X}\boldsymbol{\beta})|$. For each simulation setup, all summary statistics are computed based on a total of 200 Monte-Carlo data sets. The numerical results are summarized in Tables 1–6.

From Tables 1–4, several observations can be made as follows. First, in model (3.1) where the true relationship is linear, the standard variable selection methods perform slightly better than their distribution-transformation counterparts. This is expected because the new proposed methods are by no means tailored to a specific parametric model. Nevertheless, our methods are still competitive in this case. In other models, as we can see, the distribution-transformation penalized methods are clearly the winners. This is more evident in models (3.3) and (3.4) where the standard methods break down with intolerably low true positive rate. Because the identity function plays the role of the transformation in the standard methods, it also sheds some light on the advantage of adopting a bounded (distribution) transformation. Second, a simple comparison of results from Tables 1 and 2, or Tables 3 and 4, implies that our methods are very robust to the violation of the linearity condition. This is in accordance with our discussions in Section 2.1. Third, we then focus on the performance of our non-convex penalized estimation methods. Generally, the D-SCAD and the D-MCP perform well in terms of high estimation accuracy: the reported values of COR_1 and COR_2 are very close to 1. In contrast, the D-LASSO is less appealing which suggests that it may not have the oracle property. Furthermore, we can observe that all the values of TPR are very close to 1: the lowest is 90.50%. This means that, for each estimator, significant predictors can be identified with desired high probability. However, as compared with the D-SCAD and D-MCP, the D-LASSO is too greedy in the sense that it over-selects many insignificant predictors. The FDR values are alarmingly high: the lowest is 77.08%, and the highest is 81.46%. The numerical results on MS and FDR further indicate that the performance of the D-SCAD lies between those of the other two. This kind of phenomenon was also observed in [5]. Finally, from Tables 1, 3, 5 and 6, it can also be seen that the estimation performance of those three estimators deteriorates as the predictor dimension increases from $p = 500$ to $p = 2000$, and for the D-SCAD and the D-MCP, the TPR values become smaller as the correlation among the predictors gets larger.

Example 7. We now apply the proposed methods to the Colon gene expression data which were previously analyzed in [1]. This benchmark data set contains the expression of 2000 genes in 22 normal tissues and 40 colon tumor tissues, and is available at http://stat.ethz.ch/~dettling/bagboost.html. As done by Dudoit et al. [14], we standardize each sample to have zero mean and unit variance, and perform the study by randomly splitting the 62 samples into training and test sets. Specifically, we set two-thirds of the observations from the tumor class and two-thirds of the observations from the normal class as training samples, and the rest as test samples. The D-SCAD and the D-MCP are then applied to the training data. As before, the tuning parameters λ and a are selected by tenfold cross-validation. For each sparse index estimator $\hat{\boldsymbol{\beta}}$, we fit a

Table 2
Simulation results for Example 2.

Selection method	MS (mean)	TPR (mean)	FDR (mean)	COR_1 (mean ± sd)	COR_2 (mean ± sd)
Model (3.1)					
D-SCAD	21.635	0.994	0.7213	0.9549 ± 0.0388	0.9624 ± 0.0342
D-MCP	7.205	0.972	0.2199	0.9644 ± 0.0367	0.9703 ± 0.0319
D-LASSO	33.435	0.985	0.8035	0.9044 ± 0.0590	0.9183 ± 0.0548
SCAD	19.850	0.994	0.6817	0.9652 ± 0.0401	0.9716 ± 0.0324
MCP	7.665	0.984	0.2316	0.9722 ± 0.0289	0.9773 ± 0.0239
LASSO	32.010	0.993	0.7939	0.9221 ± 0.0501	0.9345 ± 0.0451
Model (3.2)					
D-SCAD	8.895	1	0.2542	0.9977 ± 0.0021	0.9982 ± 0.0016
D-MCP	6.340	1	0.1335	0.9972 ± 0.0028	0.9978 ± 0.0024
D-LASSO	31.885	1	0.7987	0.9921 ± 0.0061	0.9937 ± 0.0050
SCAD	12.575	1	0.3836	0.9942 ± 0.0061	0.9954 ± 0.0050
MCP	6.785	1	0.1632	0.9937 ± 0.0077	0.9950 ± 0.0063
LASSO	34.340	1	0.8231	0.9825 ± 0.0112	0.9856 ± 0.0096
Model (3.3)					
D-SCAD	18.885	1.000	0.6423	0.9774 ± 0.0211	0.9817 ± 0.0173
D-MCP	7.765	0.994	0.2332	0.9801 ± 0.0210	0.9837 ± 0.0180
D-LASSO	34.825	0.999	0.8089	0.9408 ± 0.0394	0.9509 ± 0.0341
SCAD	16.440	0.641	0.6920	0.6373 ± 0.1711	0.6733 ± 0.1751
MCP	5.820	0.478	0.4082	0.6093 ± 0.2349	0.6439 ± 0.2424
LASSO	17.570	0.612	0.6313	0.6023 ± 0.1967	0.6381 ± 0.2041
Model (3.4)					
D-SCAD	14.610	1.000	0.4966	0.9876 ± 0.0125	0.9901 ± 0.0103
D-MCP	6.985	0.997	0.1839	0.9898 ± 0.0119	0.9918 ± 0.0108
D-LASSO	34.195	0.999	0.8146	0.9602 ± 0.0244	0.9670 ± 0.0212
SCAD	10.045	0.393	0.5304	0.4105 ± 0.2817	0.4394 ± 0.2981
MCP	3.410	0.250	0.3157	0.3593 ± 0.3123	0.3873 ± 0.3265
LASSO	11.660	0.374	0.5086	0.3955 ± 0.2762	0.4202 ± 0.2916

Table 3
Simulation results for Example 3.

Selection method	MS (mean)	TPR (mean)	FDR (mean)	COR_1 (mean ± sd)	COR_2 (mean ± sd)
Model (3.1)					
D-SCAD	22.160	0.975	0.7457	0.9470 ± 0.0377	0.9438 ± 0.0404
D-MCP	8.605	0.923	0.3441	0.9478 ± 0.0459	0.9446 ± 0.0496
D-LASSO	31.170	0.989	0.7988	0.9247 ± 0.0444	0.9186 ± 0.0497
SCAD	21.995	0.990	0.7327	0.9649 ± 0.0264	0.9623 ± 0.0286
MCP	8.335	0.963	0.3040	0.9667 ± 0.0287	0.9643 ± 0.0308
LASSO	31.750	0.996	0.7950	0.9408 ± 0.0307	0.9364 ± 0.0337
Model (3.2)					
D-SCAD	9.455	1	0.2738	0.9966 ± 0.0024	0.9964 ± 0.0026
D-MCP	6.365	1	0.1269	0.9963 ± 0.0029	0.9961 ± 0.0031
D-LASSO	28.645	1	0.7708	0.9904 ± 0.0060	0.9897 ± 0.0069
SCAD	14.955	0.997	0.5350	0.9893 ± 0.0119	0.9884 ± 0.0135
MCP	7.240	0.994	0.2036	0.9906 ± 0.0108	0.9899 ± 0.0118
LASSO	32.490	1.000	0.8135	0.9730 ± 0.0152	0.9703 ± 0.0181
Model (3.3)					
D-SCAD	20.950	0.982	0.6925	0.9623 ± 0.0380	0.9593 ± 0.0418
D-MCP	7.725	0.939	0.2752	0.9660 ± 0.0410	0.9634 ± 0.0451
D-LASSO	31.000	0.997	0.7815	0.9411 ± 0.0390	0.9364 ± 0.0431
SCAD	8.525	0.338	0.5459	0.4687 ± 0.2952	0.4544 ± 0.2871
MCP	2.795	0.218	0.3067	0.4137 ± 0.3204	0.3978 ± 0.3104
LASSO	10.170	0.343	0.5032	0.4393 ± 0.3080	0.4241 ± 0.3017
Model (3.4)					
D-SCAD	16.980	0.997	0.6053	0.9842 ± 0.0165	0.9830 ± 0.0175
D-MCP	7.395	0.993	0.2172	0.9864 ± 0.0146	0.9855 ± 0.0162
D-LASSO	31.335	0.999	0.7992	0.9646 ± 0.0225	0.9612 ± 0.0266
SCAD	6.135	0.255	0.4754	0.3772 ± 0.2813	0.3584 ± 0.2703
MCP	1.850	0.155	0.2546	0.3211 ± 0.3075	0.3055 ± 0.2924
LASSO	7.250	0.245	0.4226	0.3488 ± 0.2918	0.3308 ± 0.2809

Table 4
Simulation results for Example 4.

Selection method	MS (mean)	TPR (mean)	FDR (mean)	COR_1 (mean ± sd)	COR_2 (mean ± sd)
Model (3.1)					
D-SCAD	21.640	0.962	0.7365	0.9481 ± 0.0413	0.9541 ± 0.0353
D-MCP	8.080	0.905	0.3040	0.9503 ± 0.0466	0.9557 ± 0.0399
D-LASSO	31.480	0.987	0.7951	0.9280 ± 0.0384	0.9359 ± 0.0370
SCAD	20.760	0.988	0.7070	0.9641 ± 0.0311	0.9675 ± 0.0275
MCP	7.925	0.949	0.2779	0.9666 ± 0.0347	0.9696 ± 0.0311
LASSO	30.195	0.993	0.7920	0.9415 ± 0.0369	0.9480 ± 0.0339
Model (3.2)					
D-SCAD	9.655	1	0.2871	0.9975 ± 0.0020	0.9977 ± 0.0019
D-MCP	6.600	1	0.1510	0.9972 ± 0.0023	0.9975 ± 0.0021
D-LASSO	30.800	1	0.7912	0.9920 ± 0.0051	0.9931 ± 0.0045
SCAD	13.860	1.000	0.4736	0.9923 ± 0.0082	0.9930 ± 0.0078
MCP	6.685	1.000	0.1655	0.9928 ± 0.0071	0.9935 ± 0.0063
LASSO	33.250	1.000	0.8116	0.9804 ± 0.0123	0.9828 ± 0.0113
Model (3.3)					
D-SCAD	19.205	0.993	0.6742	0.9719 ± 0.0289	0.9744 ± 0.0250
D-MCP	7.945	0.984	0.2707	0.9762 ± 0.0263	0.9782 ± 0.0234
D-LASSO	30.550	0.998	0.7898	0.9505 ± 0.0269	0.9560 ± 0.0253
SCAD	13.195	0.521	0.6564	0.6091 ± 0.2198	0.6284 ± 0.2239
MCP	4.180	0.325	0.4173	0.5554 ± 0.2749	0.5648 ± 0.2823
LASSO	14.100	0.504	0.5983	0.5737 ± 0.2553	0.5891 ± 0.2622
Model (3.4)					
D-SCAD	15.500	0.999	0.5473	0.9873 ± 0.0115	0.9883 ± 0.0108
D-MCP	7.115	0.995	0.1858	0.9886 ± 0.0109	0.9895 ± 0.0102
D-LASSO	30.265	1.000	0.7862	0.9677 ± 0.0172	0.9713 ± 0.0167
SCAD	8.545	0.347	0.5344	0.4471 ± 0.2863	0.4645 ± 0.2935
MCP	3.105	0.210	0.3323	0.3800 ± 0.3130	0.3884 ± 0.3187
LASSO	8.475	0.329	0.4627	0.4290 ± 0.2990	0.4414 ± 0.3086

Table 5
Simulation results for Example 5.

Selection method	MS (mean)	TPR (mean)	FDR (mean)	COR_1 (mean ± sd)	COR_2 (mean ± sd)
Model (3.1)					
D-SCAD	33.655	0.978	0.8386	0.9185 ± 0.0666	0.9142 ± 0.0749
D-MCP	9.160	0.917	0.3515	0.9231 ± 0.0759	0.9203 ± 0.0809
D-LASSO	42.700	0.952	0.8299	0.8470 ± 0.0809	0.8366 ± 0.0918
Model (3.2)					
D-SCAD	13.840	1	0.3845	0.9952 ± 0.0037	0.9949 ± 0.0040
D-MCP	6.720	1	0.1519	0.9952 ± 0.0039	0.9948 ± 0.0043
D-LASSO	41.380	1	0.8405	0.9861 ± 0.0086	0.9851 ± 0.0097
Model (3.3)					
D-SCAD	29.305	0.995	0.7789	0.9602 ± 0.0413	0.9582 ± 0.0442
D-MCP	8.285	0.977	0.2708	0.9682 ± 0.0328	0.9665 ± 0.0357
D-LASSO	44.370	0.985	0.8476	0.9054 ± 0.0580	0.8997 ± 0.0648
Model (3.4)					
D-SCAD	24.680	1.000	0.7084	0.9806 ± 0.0167	0.9796 ± 0.0175
D-MCP	7.415	0.997	0.2056	0.9850 ± 0.0149	0.9838 ± 0.0164
D-LASSO	47.245	0.997	0.8598	0.9355 ± 0.0378	0.9296 ± 0.0437

logistic regression model with the linear predictor $\mathbf{X}\hat{\boldsymbol{\beta}}$ as the input. The performance is evaluated by the test samples. In addition, we compare our methods with the L_1-penalized logistic regression method (PLR, [32]) and the nearest shrunken centroids method (NSC, [37]). The implementation of the NSC is accessible from the public domain *R* language package *pamr*. To reduce the variability, we repeat the above procedures 100 times, and summarize the results in Table 7. We can see that our methods are very competitive to the NSC: the classification errors of the D-SCAD and the D-MCP are not far from the NSC whereas, the NSC is not good in predictor selection because the selected number of genes is very large. The D-SCAD and L_1 PLR have the same performance, which is slightly better than the D-MCP. It is noteworthy that the maximum likelihood estimates for generalized linear models still work under the link violation [29]. Nevertheless, how to derive the theoretical properties of penalized likelihood estimation, in high-dimensional generalized linear models and under the link violation, is an interesting topic for our future study.

Table 6
Simulation results for Example 6.

Selection method	MS (mean)	TPR (mean)	FDR (mean)	COR_1 (mean ± sd)	COR_2 (mean ± sd)
Model (3.1)					
D-SCAD	30.445	0.922	0.8162	0.9166 ± 0.0577	0.9382 ± 0.0388
D-MCP	8.735	0.835	0.3966	0.9200 ± 0.0611	0.9409 ± 0.0424
D-LASSO	38.860	0.951	0.8195	0.8953 ± 0.0538	0.9223 ± 0.0440
Model (3.2)					
D-SCAD	14.065	1	0.3732	0.9958 ± 0.0039	0.9967 ± 0.0034
D-MCP	6.985	1	0.1630	0.9959 ± 0.0036	0.9968 ± 0.0031
D-LASSO	39.725	1	0.8386	0.9875 ± 0.0082	0.9904 ± 0.0066
Model (3.3)					
D-SCAD	30.030	0.961	0.8035	0.9463 ± 0.0473	0.9583 ± 0.0350
D-MCP	8.475	0.903	0.3225	0.9508 ± 0.0556	0.9627 ± 0.0400
D-LASSO	40.565	0.981	0.8299	0.9235 ± 0.0411	0.9440 ± 0.0316
Model (3.4)					
D-SCAD	26.435	0.987	0.7482	0.9767 ± 0.0283	0.9813 ± 0.0212
D-MCP	7.590	0.974	0.2317	0.9808 ± 0.0272	0.9847 ± 0.0205
D-LASSO	42.270	0.994	0.8386	0.9518 ± 0.0266	0.9642 ± 0.0200

Table 7
Colon data: classification errors for the four methods over 100 random splits into training and test sets of the 62 samples.

Method	Training error (median)	Test error (median)	No. of selected genes (median)
PLR	2/41	3/21	10
NSC	4/41	3/21	87
D-SCAD	2/41	3/21	10
D-MCP	4/41	4/21	2

Appendix. Proofs

Proof of Proposition 1. This follows immediately from Theorem 2.1 in [29]. □

Proof of Theorem 1. Let $\mathbf{C}_0 = n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T$, $\mathbf{c}_0 = n^{-1}\sum_{i=1}^n h_i\mathbf{x}_i$, $\bar{\mathbf{x}} = n^{-1}\sum_{i=1}^n \mathbf{x}_i$ and $\bar{h} = n^{-1}\sum_{i=1}^n h_i$, where $h_i = h(y_i)$. Let $\mathbf{C} = \mathbf{C}_0 - \bar{\mathbf{x}}\bar{\mathbf{x}}^T$ and $\mathbf{c} = \mathbf{c}_0 - \bar{h}\bar{\mathbf{x}}$. Then, $\hat{\boldsymbol{\beta}}_h = \mathbf{C}^{-1}\mathbf{c}$. A simple calculation yields that

$$\begin{aligned}\mathbf{C}^{-1}\mathbf{c} - \boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma} &= \boldsymbol{\Sigma}^{-1}(\mathbf{c}-\boldsymbol{\sigma}) + (\mathbf{C}^{-1}-\boldsymbol{\Sigma}^{-1})\boldsymbol{\sigma} + (\mathbf{C}^{-1}-\boldsymbol{\Sigma}^{-1})(\mathbf{c}-\boldsymbol{\sigma})\\ &\equiv T_1 + T_2 + T_3. \end{aligned} \tag{A.5}$$

First, we consider T_1. By condition (A1), we have

$$E(\|\mathbf{c}_0-\boldsymbol{\sigma}\|_2^2) = \frac{1}{n}E\{(h_1\mathbf{x}_1-\boldsymbol{\sigma})^T(h_1\mathbf{x}_1-\boldsymbol{\sigma})\} \le \frac{1}{n}E(h_1^2\mathbf{x}_1^T\mathbf{x}_1) = O\left(\frac{p}{n}\right).$$

It follows that

$$\|\mathbf{c}_0-\boldsymbol{\sigma}\|_2 = O_P\left(\sqrt{\frac{p}{n}}\right). \tag{A.6}$$

Similarly, we obtain

$$\|\bar{\mathbf{x}}\|_2 = O_P\left(\sqrt{\frac{p}{n}}\right). \tag{A.7}$$

By (A.6) and (A.7),

$$\|\mathbf{c}-\boldsymbol{\sigma}\|_2 \le \|\mathbf{c}_0-\boldsymbol{\sigma}\|_2 + \|\bar{h}\bar{\mathbf{x}}\|_2 = O_P\left(\sqrt{\frac{p}{n}}\right). \tag{A.8}$$

Thus, by condition (A1), we have

$$\|T_1\|_2 = O_P\left(\sqrt{\frac{p}{n}}\right). \tag{A.9}$$

Next, we consider T_2. Note that

$$\mathbf{C}^{-1} - \boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \mathbf{C})\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \mathbf{C})(\mathbf{C}^{-1} - \boldsymbol{\Sigma}^{-1}). \tag{A.10}$$

By the Cauchy–Schwarz inequality and condition (A2), we have

$$E(\|\mathbf{C}_0 - \boldsymbol{\Sigma}\|_2^2) \le \frac{p}{n} E\left(\sum_{j=1}^{p} x_{1j}^4\right) = O\left(\frac{p^2}{n}\right).$$

It implies that

$$\|\mathbf{C}_0 - \boldsymbol{\Sigma}\|_2 = O_P\left(\frac{p}{\sqrt{n}}\right). \tag{A.11}$$

By (A.7) and (A.11),

$$\|\mathbf{C} - \boldsymbol{\Sigma}\|_2 \le \|\mathbf{C}_0 - \boldsymbol{\Sigma}\|_2 + \|\bar{\mathbf{x}}\bar{\mathbf{x}}^T\|_2 = O_P\left(\frac{p}{\sqrt{n}}\right). \tag{A.12}$$

This, together with (A.10), yields that

$$\|\mathbf{C}^{-1} - \boldsymbol{\Sigma}^{-1}\|_2 = O_P\left(\frac{p}{\sqrt{n}}\right). \tag{A.13}$$

By condition (A1), we have

$$\begin{aligned}\|T_2\|_2^2 &= \operatorname{trace}\{(\mathbf{C}^{-1} - \boldsymbol{\Sigma}^{-1})\boldsymbol{\sigma}\boldsymbol{\sigma}^T(\mathbf{C}^{-1} - \boldsymbol{\Sigma}^{-1})^T\} \\ &\le \operatorname{trace}\{(\mathbf{C}^{-1} - \boldsymbol{\Sigma}^{-1})E(h_1^2\mathbf{x}_1\mathbf{x}_1^T)(\mathbf{C}^{-1} - \boldsymbol{\Sigma}^{-1})^T\} \\ &= O_P(\|\mathbf{C}^{-1} - \boldsymbol{\Sigma}^{-1}\|_2^2).\end{aligned}$$

Then, we obtain

$$\|T_2\|_2 = O_P\left(\frac{p}{\sqrt{n}}\right). \tag{A.14}$$

As for the term T_3, by (A.8) and (A.13),

$$\|T_3\|_2 = O_P\left(\sqrt{\frac{p^3}{n^2}}\right). \tag{A.15}$$

Therefore, by (A.9), (A.14), (A.15) and condition (A3), we have

$$\|\mathbf{C}^{-1}\mathbf{c} - \boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}\|_2 = O_P\left(\frac{p}{\sqrt{n}}\right) = o_P(1). \tag{A.16}$$

The proof is complete. □

Proof of Theorem 2. Write $\mathbf{C}^{-1}\mathbf{c} - \boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma} = M + R$, where

$$\begin{aligned}M &= \boldsymbol{\Sigma}^{-1}(\mathbf{c}_0 - \boldsymbol{\sigma}) + \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \mathbf{C}_0)\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma} - \bar{h}\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}} \\ &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Sigma}^{-1}(h_i - \mathbf{x}_i^T\boldsymbol{\beta}_h)\mathbf{x}_i - \bar{h}\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}} \\ &\equiv M_1 - M_2.\end{aligned}$$

First, we derive the asymptotic distribution of $\sqrt{n}\mathbf{v}^T M$. Consider M_2. Note that

$$M_2 = \frac{1}{n^2}\sum_{i=1}^{n} h_i\boldsymbol{\Sigma}^{-1}\mathbf{x}_i + \frac{n-1}{n}U = \frac{n-1}{n}U + O_P\left(\frac{\sqrt{p}}{n}\right),$$

where U is a standard U-statistic with $(h_i\boldsymbol{\Sigma}^{-1}\mathbf{x}_j + h_j\boldsymbol{\Sigma}^{-1}\mathbf{x}_i)/2$ as the kernel. Let $\hat{U}$ be the projection of the U-statistic U. A simple calculation yields that

$$\hat{U} = \frac{1}{n}\sum_{i=1}^{n} E(h_1)\boldsymbol{\Sigma}^{-1}\mathbf{x}_i.$$

Then, by Theorem 5.3.2 in [35], we obtain

$$U = \hat{U} + O_P\left(\frac{\sqrt{p}}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} E(h_1)\boldsymbol{\Sigma}^{-1}\mathbf{x}_i + O_P\left(\frac{\sqrt{p}}{n}\right).$$

It follows that

$$M_2 = \frac{1}{n}\sum_{i=1}^{n} E(h_1)\boldsymbol{\Sigma}^{-1}\mathbf{x}_i + O_P\left(\frac{\sqrt{p}}{n}\right).$$

Thus, we have $\sqrt{n}\mathbf{v}^T M = \sum_{i=1}^{n}\mathbf{z}_{ni} + o_P(1)$, where

$$\mathbf{z}_{ni} = \frac{1}{\sqrt{n}}\mathbf{v}^T\boldsymbol{\Sigma}^{-1}\{\tilde{h}_i - \mathbf{x}_i\boldsymbol{\beta}_h\}\mathbf{x}_i \quad \text{for } i = 1, \ldots, n.$$

Because $\Lambda = \mathrm{Cov}[\boldsymbol{\Sigma}^{-1}\{\tilde{h}(Y) - X^T\boldsymbol{\beta}_h\}X]$ by definition, as $\mathbf{v}^T\Lambda\mathbf{v} \to G$, we obtain

$$\sum_{i=1}^{n}\mathrm{Cov}(\mathbf{z}_{ni}) = n\mathrm{Cov}(\mathbf{z}_{n1}) \to G. \tag{A.17}$$

On the other hand, for any $\varepsilon > 0$,

$$\begin{aligned}\sum_{i=1}^{n} E[\|\mathbf{z}_{ni}\|_2^2\mathbf{1}\{\|\mathbf{z}_{ni}\|_2 > \varepsilon\}] &= nE[\|\mathbf{z}_{n1}\|_2^2\mathbf{1}\{\|\mathbf{z}_{n1}\|_2 > \varepsilon\}] \\ &\le n(E\|\mathbf{z}_{n1}\|_2^4)^{1/2}\{P(\|\mathbf{z}_{n1}\|_2 > \varepsilon)\}^{1/2}.\end{aligned} \tag{A.18}$$

Because $\mathbf{v}^T\Lambda\mathbf{v} \to G > 0$ as $n \to \infty$, we have

$$P(\|\mathbf{z}_{n1}\|_2 > \varepsilon) \le \frac{E(\|\mathbf{z}_{n1}\|_2^2)}{\varepsilon^2} = O\left(\frac{1}{n}\right). \tag{A.19}$$

A simple calculation yields that,

$$\begin{aligned}\|\mathbf{z}_{n1}\|_2^2 &= \frac{1}{n}\|\mathbf{v}^T\boldsymbol{\Sigma}^{-1}\{\tilde{h}_1 - \mathbf{x}_1^T\boldsymbol{\beta}_h\}\mathbf{x}_1\|_2^2 \\ &\le \|\boldsymbol{\Sigma}^{-1}\mathbf{v}\|_2^2\frac{1}{n}\|\{\tilde{h}_1 - \mathbf{x}_1^T\boldsymbol{\beta}_h\}\mathbf{x}_1\|_2^2 \\ &\le \frac{2}{L_1^2 n}(\tilde{h}_1^2\mathbf{x}_1^T\mathbf{x}_1 + \boldsymbol{\beta}_h^T\mathbf{x}_1\mathbf{x}_1^T\mathbf{x}_1\mathbf{x}_1^T\boldsymbol{\beta}_h) \\ &\le \sup_y h^2(y)\frac{4}{L_1^2 n}\left\{\mathbf{x}_1^T\mathbf{x}_1 + L_2(\mathbf{x}_1^T\mathbf{x}_1)^2\right\}.\end{aligned}$$

Because $\max_{1\le j\le p} E(X_j^8) < L_4 < \infty$, it follows that

$$E(\|\mathbf{z}_{n1}\|_2^4) = O\left(\frac{p^3}{n^2}\right). \tag{A.20}$$

By (A.18)–(A.20) and $p = o(n^{1/4})$, we have

$$\sum_{i=1}^{n} E\|\mathbf{z}_{ni}\|_2^2\mathbf{1}\{\|\mathbf{z}_{ni}\|_2 > \varepsilon\} = nO\left(\sqrt{\frac{p^3}{n^2}}\right)O\left(\frac{1}{\sqrt{n}}\right) = o(1). \tag{A.21}$$

Therefore, $\mathbf{z}_{ni}$ satisfies the conditions of the Lindeberg–Feller central limit theorem. This means that $\sqrt{n}\mathbf{v}^T M$ has an asymptotic normal distribution $N(0, G)$.

It remains to prove that $\sqrt{n}\mathbf{v}^T R = o_P(1)$, where

$$\begin{aligned}R &= \boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}\bar{\mathbf{x}}^T\boldsymbol{\beta}_h + \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \mathbf{C})(\mathbf{C}^{-1} - \boldsymbol{\Sigma}^{-1})\boldsymbol{\sigma} + (\mathbf{C}^{-1} - \boldsymbol{\Sigma}^{-1})(\mathbf{c} - \boldsymbol{\sigma}) \\ &\equiv R_1 + R_2 + R_3.\end{aligned}$$

Consider the terms R_1 and R_2. It is easy to show that

$$\|R_1\|_2 = O_P\left(\frac{p}{n}\right) \tag{A.22}$$

and

$$\|R_2\|_2 = O_P\left(\frac{p^2}{n}\right). \tag{A.23}$$

Furthermore, we have

$$\|R_3\|_2 = \|T_3\|_2 = O_P\left(\sqrt{\frac{p^3}{n^2}}\right). \tag{A.24}$$

Thus, by (A.22)–(A.24) and $p = o(n^{1/4})$, we obtain

$$\sqrt{n}\mathbf{v}^T R = \sqrt{n} O_P\left(\frac{p^2}{n}\right) = o_P(1). \tag{A.25}$$

The proof is complete. □

Lemma 1. *Assume that for some $\gamma > 0$, $p = O(n^\gamma)$, and for some $\delta, M > 0$,*

$$E|X_j - E(X_j)|^{4\gamma+4+\delta} \le M \quad \text{forall } j = 1, \ldots, p.$$

Then, $\max_{1\le i,j\le p} |\mathbf{C}_{ij} - \boldsymbol{\Sigma}_{ij}| = O_P(\sqrt{\log p/n})$.

Proof of Lemma 1. The result follows from the Bernstein inequality as well as a truncation argument. Its proof can be found in [7]. □

Corollary 1. *Assume the conditions of Theorem 3. Then, we have*

$$\lim_{n\to\infty} P(\lambda_{\min}(\mathbf{C}_{(1,1)}) \ge M_1) = 1 \quad \textit{and} \quad \lim_{n\to\infty} P\left(\max_{1\le j\le p} \mathbf{C}_{jj} \le M_3^{\frac{2}{4(c_2-c_1)k+4+\delta}}\right) = 1.$$

Proof of Corollary 1. By Lemma 1, for any vector $\mathbf{v} = (v_1, \ldots, v_q)^T \in \mathbf{R}^q$ such that $\|\mathbf{v}\|_2 = 1$,

$$|\mathbf{v}^T(\mathbf{C}_{(1,1)} - \boldsymbol{\Sigma}_{(1,1)})\mathbf{v}| \le \max_{1\le i,j\le p} |\mathbf{C}_{ij} - \boldsymbol{\Sigma}_{ij}| \left(\sum_{1\le i\le q} |v_i|\right)^2 = O_P\left(q\sqrt{\frac{\log p}{n}}\right).$$

By conditions (A) and (B), we obtain $\lim_{n\to\infty} P(\lambda_{\min}(\mathbf{C}_{(1,1)}) \ge M_1) = 1$. Because

$$\max_{1\le j\le p} \mathbf{C}_{jj} \le \max_{1\le j\le p} \boldsymbol{\Sigma}_{jj} + \max_{1\le j\le p} |\mathbf{C}_{jj} - \boldsymbol{\Sigma}_{jj}|$$

and by condition (F),

$$\max_{1\le j\le p} E|X_j - E(X_j)|^2 \le M_3^{\frac{2}{4(c_2-c_1)k+4+\delta}},$$

the second limit follows immediately. The proof is complete. □

Proof of Theorem 3. Before proceeding, we note that since $q = o(n)$ under condition (B), the random matrix $\mathbf{X}_{(1)}$ has rank q with probability one, whence the matrix $\mathbf{C}_{(1,1)}$ is invertible with probability one. Define the event $E_{n,K}$ by

$$E_{n,K} = \{\lambda_{\min}(\mathbf{C}_{(1,1)}) \ge M_1 \text{ and } \mathbf{C}_{jj} \le K \text{ for all } j = 1, \ldots, p\}.$$

By Corollary 1, when K is sufficiently large, we have $P(E_{n,K}) \to 1$ as $n \to \infty$. Denote by A_n the event that $\hat{\boldsymbol{\beta}}^o$ is a local minimum of $Q_\lambda(\boldsymbol{\beta})$. Note that

$$P(A_n) = P(A_n|E_{n,K})P(E_{n,K}) + P(A_n|E_{n,K}^c)P(E_{n,K}^c) \ge P(A_n|E_{n,K})P(E_{n,K}).$$

Thus, to prove the theorem, it suffices to show that $P(A_n|E_{n,K}) \to 1$ as $n \to \infty$. The arguments below are then conditioned on the event $E_{n,K}$, for a sufficiently large K.

Let $\mathcal{P} = \{1, 2, \ldots, q\}$ and $\mathcal{N} = \{q+1, \ldots, p\}$. First, we show that for any $a > 0$,

$$P(|\hat{\beta}_j^o| > a\lambda \text{ for all } j \in \mathcal{P}) \to 1 \quad \text{as } n \to \infty. \tag{A.26}$$

Because $|\hat{\beta}_j^o| \ge |\beta_{0,j}| - |\hat{\beta}_j^o - \beta_{0,j}|$, by condition (C) and $\lambda = o(n^{-(1-c_2+c_1)/2})$, it suffices to show that

$$\max_{j\in\mathcal{P}} |\hat{\beta}_j^o - \beta_{0,j}| = o_P\left(n^{-\frac{1-c_2}{2}}\right), \tag{A.27}$$

or equivalently

$$\max_{j\in\mathcal{P}}|z_j| = o_P\left(n^{\frac{c_2}{2}}\right), \tag{A.28}$$

where $z_j = \sqrt{n}(\hat{\beta}_j^o - \beta_{0,j})$.

Let $\mathbf{z} = (z_1, \ldots, z_q)^T$ and $\mathbf{e} = (e_1, \ldots, e_n)^T = \mathbf{F}_n(\mathbf{y}) - \mathbf{X}_{(1)}^T\boldsymbol{\beta}_{(1)}$. Write

$$\hat{\boldsymbol{\beta}}_{(1)}^o = (\mathbf{C}_{(1,1)})^{-1}\frac{1}{n}\mathbf{X}_{(1)}^T\mathbf{F}_n(\mathbf{y}) = \boldsymbol{\beta}_{(1)} + (\mathbf{C}_{(1,1)})^{-1}\frac{1}{n}\mathbf{X}_{(1)}^T\mathbf{e}$$

and

$$\mathbf{z} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{(1)}^o - \boldsymbol{\beta}_{(1)}) = (\mathbf{C}_{(1,1)})^{-1}\frac{1}{\sqrt{n}}\mathbf{X}_{(1)}^T\mathbf{e} \equiv \mathbf{H}_{(1)}^T\mathbf{e}, \tag{A.29}$$

where $\mathbf{H}_{(1)}^T = (\mathbf{h}_{(1),1}, \ldots, \mathbf{h}_{(1),q})^T = n^{-1/2}(\mathbf{C}_{(1,1)})^{-1}\mathbf{X}_{(1)}^T$. Because $\mathbf{H}_{(1)}^T\mathbf{H}_{(1)} = (\mathbf{C}_{(1,1)})^{-1}$, we have $\|\mathbf{h}_{(1),j}\|_2^2 \le M_1^{-1}$ for all $j \in \mathcal{P}$ on the event $E_{n,K}$.

Let $\mathbf{e}^* = (e_1^*, \ldots, e_n^*)^T = \mathbf{F}(\mathbf{y}) - \mathbf{X}_{(1)}^T\boldsymbol{\beta}_{(1)}$ and $\mathbf{z}^* = (z_1^*, \ldots, z_q^*)^T = \mathbf{H}_{(1)}^T\mathbf{e}^*$. Note that $E(e_i^*)^{2k} < \infty$ under condition (D). Then, an application of Lemma 3 in [38] yields that $E(z_j^*)^{2k} < \infty$ for all $j \in \mathcal{P}$. It follows that $P(|z_j^*| > t) = O(t^{-2k})$ for any $t > 0$. By condition (B), for any $\eta > 0$, we obtain

$$\begin{aligned} P\left(|z_j| > \eta n^{\frac{c_2}{2}} \text{ for some } j \in \mathcal{P}\right) &\le \sum_{j=1}^{q} P\left(|z_j^*| > \eta n^{\frac{c_2}{2}} - |z_j - z_j^*|\right) \\ &\le \sum_{j=1}^{q}\frac{1}{\eta^{2k}}n^{-c_2k} \\ &= \frac{1}{\eta^{2k}}qn^{-c_2k} = O(n^{-(c_2-c_1)k}) = o(1), \end{aligned}$$

where in the second inequality we use the fact that $\sup_y |F_n(y) - F(y)| = o(\log n/\sqrt{n})$, so that $\max_{j\in\mathcal{P}}|z_j - z_j^*| = o_P(n^{c_2/2})$.

Let $S_j(\boldsymbol{\beta}) = n^{-1}\mathbf{X}_j^T\{\mathbf{F}_n(\mathbf{y}) - \mathbf{X}\boldsymbol{\beta}\}$. Next, we show that as $n \to \infty$,

$$P(|S_j(\hat{\boldsymbol{\beta}}^o)| > \lambda \text{ for some } j \in \mathcal{N}) \to 0. \tag{A.30}$$

For each $j \in \mathcal{N}$, define $\xi_j = \sqrt{n}S_j(\hat{\boldsymbol{\beta}}^o)$. Note that

$$\begin{aligned} (\xi_j, j \in \mathcal{N}) &= -\frac{1}{\sqrt{n}}\mathbf{X}_{(2)}^T\{\mathbf{F}_n(\mathbf{y}) - \mathbf{X}_{(1)}\hat{\boldsymbol{\beta}}_{(1)}^o\} \\ &= -\frac{1}{\sqrt{n}}\mathbf{X}_{(2)}^T\left\{\mathbf{I} - \frac{1}{n}\mathbf{X}_{(1)}(\mathbf{C}_{(1,1)})^{-1}\mathbf{X}_{(1)}^T\right\}\mathbf{e} \\ &\equiv \mathbf{H}_{(2)}^T\mathbf{e}. \end{aligned}$$

Thus, we have $\xi_j = \mathbf{h}_{(2),j}^T\mathbf{e}$ with $\mathbf{h}_{(2),j}$ being the jth column vector of $\mathbf{H}_{(2)}$. Further, a direct calculation yields that

$$\mathbf{H}_{(2)}^T\mathbf{H}_{(2)} = \frac{1}{n}\mathbf{X}_{(2)}^T\left\{\mathbf{I} - \frac{1}{n}\mathbf{X}_{(1)}(\mathbf{C}_{(1,1)})^{-1}\mathbf{X}_{(1)}^T\right\}\mathbf{X}_{(2)}.$$

Since $\mathbf{I} - n^{-1}\mathbf{X}_{(1)}(\mathbf{C}_{(1,1)})^{-1}\mathbf{X}_{(1)}^T$ has eigenvalues between 0 and 1, we have $\|\mathbf{h}_{(2),j}\|_2^2 \le K$ for all $j \in \mathcal{N}$ on the event $E_{n,K}$. Let $\xi_j^* = \mathbf{h}_{(2),j}^T\mathbf{e}^*$. It then follows that $E(\xi_j^*)^{2k} < \infty$ and $P(|\xi_j^*| > t) = O(t^{-2k})$ for any $t > 0$. Thus, we obtain

$$\begin{aligned} P(|S_j(\hat{\boldsymbol{\beta}}^o)| > \lambda \text{ for some } j \in \mathcal{N}) &\le P(|\xi_j^*| > \sqrt{n}\lambda - |\xi_j - \xi_j^*| \text{ for some } j \in \mathcal{N}) \\ &\le \sum_{j=q+1}^{p} P(|\xi_j^*| > \sqrt{n}\lambda) \\ &= (p-q)O\left(\frac{1}{\lambda^{2k}n^k}\right) = O\left(\frac{p}{\lambda^{2k}n^k}\right) = o(1), \end{aligned}$$

where the second inequality is also due to $\sup_y |F_n(y) - F(y)| = o(\log n/\sqrt{n})$.

Finally, we show that $P(A_n|E_{n,K}) \to 1$ as $n \to \infty$. Let $\rho(t) = \rho(t;\lambda) = \lambda^{-1}J_\lambda(t)$ and $\bar\rho(t) = \text{sign}(t)\rho'(|t|)$. Following [40], we define the local concavity of the penalty function $\rho(\cdot)$ at $\mathbf{v} = (v_1,\ldots,v_q)^T \in \mathbf{R}^q$ with $\|\mathbf{v}\|_0 = q$ as

$$\kappa(\rho;\mathbf{v}) = \lim_{\epsilon\to 0+}\max_{j\in\mathcal{P}}\sup_{t_1,t_2\in(|v_j|-\varepsilon,|v_j|+\epsilon),t_1<t_2} -\frac{\rho'(t_2)-\rho'(t_1)}{t_2-t_1}. \tag{A.31}$$

For the SCAD penalty and the minimax concave penalty, by (A.26), we have in probability $\boldsymbol{\rho}(\hat{\boldsymbol{\beta}}^o_{(1)}) = (\bar\rho(\hat\beta^o_1),\ldots,\bar\rho(\hat\beta^o_q))^T = \mathbf{0}$ and $\kappa(\rho;\hat{\boldsymbol{\beta}}^o_{(1)}) = 0$. This, together with (A.30), yields that with probability approaching 1,

$$\mathbf{X}^T_{(1)}\mathbf{F}_n(\mathbf{y}) - \mathbf{X}^T_{(1)}\mathbf{X}_{(1)}\hat{\boldsymbol{\beta}}^o_{(1)} = \mathbf{0} = \boldsymbol{\rho}(\hat{\boldsymbol{\beta}}^o_{(1)}), \tag{A.32}$$

$$\max_{j\in\mathcal{N}}\left|\frac{1}{n\lambda}\mathbf{X}^T_j\{\mathbf{F}_n(\mathbf{y}) - X_{(1)}\hat{\boldsymbol{\beta}}^o_{(1)}\}\right| \le 1 = \rho'(0+), \tag{A.33}$$

$$\lambda_{\min}(\mathbf{C}_{(1,1)}) > 0 = \lambda\kappa(\rho;\hat{\boldsymbol{\beta}}^o_{(1)}). \tag{A.34}$$

Thus, the assumed conditions (18)–(20) of Theorem 1 in [30] are satisfied with probability approaching 1. This implies that $\hat{\boldsymbol{\beta}}^o$ is in probability a strict local minimizer of $Q_\lambda(\boldsymbol{\beta})$. The proof is complete. □

Proof of Theorem 4. By Theorem 1 in [34], $\lambda_{\min}(\mathbf{C}_{(1,1)}) > M_1$ holds in probability for some $M_1 > 0$. Thus, condition (B) may be relaxed to $q = O(n^{c_1})$ for some $0 \le c_1 < 1$. Under the theorem conditions, we still have $P(E_{n,K}) \to 1$ as $n \to \infty$. Note that for a Gaussian random variable W with mean 0 and variance σ^2,

$$P(|W| > t) \le \sqrt{\frac{2}{\pi}}\exp\left(-\frac{t^2}{2\sigma^2}\right) \quad \text{for } t \ge \sigma. \tag{A.35}$$

Then the theorem can be easily proved by observing that

$$P\left(|z_j| > \eta n^{\frac{c_2}{2}} \text{ for some } j \in \mathcal{P}\right) = O(qe^{n^{-c_2}}) = o(1)$$

and

$$P(|S_j(\hat{\boldsymbol{\beta}}^o)| > \lambda \text{ for some } j \in \mathcal{N}) = O(pe^{-n\lambda^2}) = o(1).$$

The proof is complete. □

References

[1] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Mack, J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proceedings of the National Academy of Sciences 96 (1999) 6745–6750.
[2] P. Alquier, G. Biau, Sparse single-index model, http://arxiv.org/abs/1101.3229v2, 2011.
[3] P.J. Bickel, Y. Ritov, A.B. Tsybakov, Simultaneous analysis of LASSO and DANTZIG selector, The Annals of Statistics 37 (2009) 1705–1732.
[4] H.D. Bondell, L. Li, Shrinkage inverse regression estimation for model-free variable selection, Journal of the Royal Statistical Society, Series B 71 (2009) 287–299.
[5] P. Breheny, J. Huang, Coordinate descent algorithms for non-convex penalized regression, with applications to biological feature selection, The Annals of Applied Statistics 5 (2011) 232–253.
[6] L. Breiman, Heuristics of instability and stabilization in model selection, The Annals of Statistics 24 (1996) 2350–2383.
[7] T. Cai, W. Liu, X. Luo, A constrained l_1 minimization approach to sparse precision matrix estimation, Journal of the American Statistical Association 106 (2011) 594–607.
[8] R.D. Cook, Regression Graphics: Ideas for Studying Regressions Through Graphics, John Wiley, New York, 1998.
[9] R.D. Cook, L. Ni, Sufficient dimension reduction via inverse regression: a minimum discrepancy approach, Journal of the American Statistical Association 100 (2005) 410–428.
[10] R.D. Cook, S. Weisberg, Discussion of sliced inverse regression for dimension reduction, by K.C. Li, Journal of the American Statistical Association 86 (1991) 316–342.
[11] A. Dalalyan, A. Juditsky, V. Spokoiny, A new algorithm for estimating the effective dimension-reduction subspace, Journal of Machine Learning Research 9 (2008) 1647–1678.
[12] A. Dalalyan, A. Tsybakov, Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity, Machine Learning 72 (2008) 39–61.
[13] A. Dalalyan, A. Tsybakov, Sparse regression learning by aggregation and Langevin Monte-Carlo, Journal of Computer and System Sciences, http://dx.doi.org/10.1016/j.jcss.2011.12.023.
[14] S. Dudoit, J. Fridlyand, T.P. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, Journal of the American Statistical Association 97 (2002) 77–87.
[15] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, The Annals of Statistics 32 (2004) 407–499.
[16] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, Journal of the American Statistical Association 96 (2001) 1348–1360.
[17] J. Fan, J. Lv, A selective overview of variable selection in high dimensional feature space, Statistica Sinica 20 (2010) 101–148.
[18] J. Fan, H. Peng, Non-concave penalized likelihood with a diverging number of parameters, The Annals of Statistics 32 (2004) 928–961.
[19] J. Friedman, T. Hastie, R. Tibshirani, Regularized paths for generalized linear models via coordinate descent, Journal of Statistical Software 33 (2010) 1–22.
[20] P. Hall, K.C. Li, On almost linearity of low dimensional projections from high dimensional data, The Annals of Statistics 21 (1993) 867–889.
[21] W. Härdle, P. Hall, H. Ichimura, Optimal smoothing in single-index models, The Annals of Statistics 21 (1993) 157–178.

[22] J.L. Horowitz, Semiparametric estimation of a regression model with an unknown transformation of the dependent variable, Econometrica 64 (1996) 103–137.
[23] M. Hristache, A. Juditsky, V. Spokoiny, Direct estimation of the index coefficient in a single-index model, The Annals of Statistics 29 (2001) 595–623.
[24] Y. Kim, H. Choi, H.S. Oh, Smoothly clipped absolute deviation on high dimensions, Journal of the American Statistical Association 103 (2008) 1665–1673.
[25] K. Knight, W.J. Fu, Asymptotics for LASSO-type estimators, The Annals of Statistics 28 (2000) 1356–1378.
[26] E. Kong, Y. Xia, Variable selection for the single-index model, Biometrika 94 (2007) 217–229.
[27] B. Li, S. Wang, On directional regression for dimension reduction, Journal of the American Statistical Association 102 (2007) 997–1008.
[28] K.C. Li, Sliced inverse regression for dimension reduction, Journal of the American Statistical Association 86 (1991) 316–327.
[29] K.C. Li, N.H. Duan, Regression analysis under link violation, The Annals of Statistics 17 (1989) 1009–1052.
[30] J. Lv, Y. Fan, A unified approach to model selection and sparse recovery using regularized least squares, The Annals of Statistics 37 (2009) 3498–3528.
[31] P. Naik, C.L. Tsai, Single-index model selections, Biometrika 88 (2001) 821–832.
[32] M.Y. Park, T. Hastie, L_1-regularization path algorithm for generalized linear models, Journal of the Royal Statistical Society, Series B 69 (2007) 659–677.
[33] S. Portnoy, Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large, I, consistency, The Annals of Statistics 12 (1984) 1298–1309.
[34] G. Raskutti, M.J. Wainwright, B. Yu, Restricted eigenvalue properties for correlated Gaussian designs, Journal of Machine Learning Research 11 (2010) 2241–2259.
[35] R.J. Serfling, Approximation Theorems of Mathematical Statistics, John Wiley & Sons, New York, 1980.
[36] R. Tibshirani, Regression shrinkage and selection via the LASSO, Journal of the Royal Statistical Society, Series B 58 (1996) 267–288.
[37] R. Tibshirani, T. Hastie, B. Narasimhan, G. Chu, Diagnosis of multiple cancer types by Shrunken centroids of gene expression, Proceedings of the National Academy of Sciences 99 (2002) 6567–6572.
[38] T. Wang, L.X. Zhu, 2011, Consistent model selection and estimation in a general single-index model with large p and small n, manuscript.
[39] Y. Wu, L. Li, Asymptotic properties of sufficient dimension reduction with a diverging number of predictors, Statistica Sinica 31 (2011) 707–730.
[40] C.H. Zhang, Nearly unbiased variable selection under minimax concave penalty, The Annals of Statistics 38 (2010) 894–942.
[41] C.H. Zhang, J. Huang, The sparsity and bias of the LASSO selection in high-dimensional linear regression, The Annals of Statistics 36 (2008) 1567–1594.
[42] P. Zhao, B. Yu, On model selection consistency of LASSO, Journal of Machine Learning Research 7 (2006) 2541–2563.
[43] L.P. Zhu, T. Wang, L.X. Zhu, L. Ferré, Sufficient dimension reduction through discretization-expectation estimation, Biometrika 97 (2010) 295–304.
[44] L.P. Zhu, L.X. Zhu, On distribution-weighted partial least squares with diverging number of highly correlated predictors, Journal of the Royal Statistical Society, Series B 71 (2009) 525–548.